

A Novel Test for Additivity in Supervised Ensemble Learners

Lucas Mentch and Giles Hooker

Department of Statistical Science
Cornell University
Ithaca, NY 14850 USA
email: 1km54@cornell.edu
gjh27@cornell.edu

November 13, 2014

Abstract

Additive models remain popular statistical tools due to their ease of interpretation and as a result, hypothesis tests for additivity have been developed to assess the appropriateness of these models. However, as data grows in size and complexity, learning algorithms continue to gain popularity due to their exceptional predictive performance. Due to the black-box nature of these learning methods, the increase in predictive power is assumed to come at the cost of interpretability and inference. However, recent work suggests that many popular learning techniques, such as bagged trees and random forests, have desirable asymptotic properties which allow for formal statistical inference when base learners are built with proper subsamples. This work extends hypothesis tests previously developed and demonstrates that by enforcing a grid structure on an appropriate test set, we may perform formal hypothesis tests for additivity among features. We develop notions of total and partial additivity and demonstrate that both tests can be carried out at no additional computational cost. We also suggest a new testing procedure based on random projections that allows for testing on larger grids, even when the grid size is larger than that of the training set. Simulations and demonstrations on real data are provided.

KEYWORDS: Bagging, Random Forest, Random Projection, Interaction, U-statistic

1 Introduction

Additive models were suggested by Friedman and Stuetzle (1981) and later generalized and made popular by Hastie and Tibshirani (1990). An underlying regression function $F: \mathcal{X} \mapsto \mathbb{R}$ is said to be additive if

$$F(X_1, \dots, X_d) = \sum_{i=1}^d F_i(X_i)$$

for some functions F_1, \dots, F_d . If the regression function cannot be written as, or at least well-approximated by, a sum of univariate functions, then an interaction exists between at least two features. Many methods have been developed to estimate the additive functions F_1, \dots, F_d including a method based on marginal integration by Linton and Nielsen (1995), a wavelet method suggested by Amato and Antoniadis (2001), and a tree-based method by Lou et al. (2013), but the most popular class of estimation methods are based on backfitting algorithms as found in Buja et al. (1989), Opsomer and Ruppert (1998, 1999), and Mammen et al. (1999).

The popularity of additive models lead naturally to hypothesis tests for additivity to asses whether an additive model is appropriate. Versions of these lack-of-fit tests have been proposed by Barry (1993), Eubank et al. (1995), Dette and Derbort (2001), Dette et al. (2001), Derbort et al. (2002), and De Canditiis and Sapatinas (2004). In some cases, such as in the work of Fan and Jiang (2005), these procedures can also evaluate whether the additive functions belong to a particular parametric class. Even when additive models are not used as the primary analytical tool, scientists often utilize tests of additivity to determine whether features (covariates) contribute additively to the response. If no interactions are detected, then levels of one feature may be changed without affecting the contribution to the response of the others.

Despite their usefulness, simple and intuitive tools like additive models can of-

ten fail to fully capture the signal hidden within modern complex data. Learning algorithms like bagged trees and random forests, both introduced by Breiman (1996, 2001), are robust to a variety of regression functions and typically generate significantly more accurate predictions. However, until recently, little was understood about the asymptotic behavior of these ensemble learners. Recently, Mentch and Hooker (2014) showed that by using subsamples instead of full bootstrap samples to build individual trees, the predictions generated by these ensemble learners can be viewed as incomplete, infinite-order U-statistics and as such, are asymptotically normal. This asymptotic normality allows for confidence intervals to accompany predictions and hypothesis tests to formally assess the significance of features. Furthermore, by imposing structure on the subsamples used in the ensemble, variance estimates may be computed at no additional cost to the original ensemble. Wager et al. (2014) develop an infinitesimal jackknife estimate of variance and Wager (2014) later showed that the conditions on subsample size imposed by Mentch and Hooker (2014) may be relaxed in the context of random forests for a specific class of tree-building methods.

This paper continues the trend of inference procedures for learning algorithms by developing formal hypothesis tests for additivity within the context of ensemble learners like bagged trees and random forests. These tests allow practitioners to formally investigate the manner in which features contribute to the response when simpler tools like additive models are insufficient. In Section 2 we propose a revised test for feature significance that imposes a grid structure on the test set and in Section 3 we demonstrate how this additional structure allows for tests of additivity. In Section 4 we extend our procedure to the case where a large test grid is needed, so as to cover the case where $p > n$. In Section 5 we provide simulations to investigate the power and α -level of our tests and in Section 6 we apply our procedure to real data.

2 An alternative test for feature significance

We begin by recalling the key result from Mentch and Hooker (2014) which paves the way for statistical inference to be carried out for ensemble learners like bagged trees and random forests. The authors demonstrate that the predictions from these ensembles may be viewed as incomplete, infinite-order U-statistics when the base learners, most often trees, are built with subsamples instead of full bootstrap samples. The kernels of these U-statistics are functions that map from the subsample to the prediction via the trees and since the result holds in the infinite order case, the size of each subsample may grow with the size of the training set, thereby allowing trees to be grown to greater depths when more data is available. Thus, given a set of subbagged trees or a subsampled random forest based on a training set of size n in which m trees are built, each with a subsample of size k , the prediction at a given point \mathbf{x}^* is asymptotically normal and the limiting variance depends only on the number of base learners relative to the size of the training set. This central limit theorem and more discussion is provided in Appendix A.

Given this asymptotic normality, Mentch and Hooker (2014) propose a hypothesis test to evaluate feature significance, which we briefly summarize here. Suppose that the data consists of only two covariates X_1 and X_2 and that the response is observed according to $Y = F(X_1, X_2) + \epsilon$. To test the significance of X_2 , we begin with a test set \mathbf{x}_{TEST} consisting of N points and build two subsampled ensembles \hat{F} and \hat{F}_1 using the same subsamples, but where \hat{F} is built using both X_1 and X_2 and \hat{F}_1 is built using only X_1 . Predictions at each point in \mathbf{x}_{TEST} are then made with each ensemble and by extension of the central limit theorem referenced above, the vector of differences in predictions has a multivariate normal distribution with mean μ and variance Σ . Given consistent estimators of these parameters, $\hat{\mu}^T \hat{\Sigma}^{-1} \hat{\mu} \sim \chi_N^2$ can be used as a test

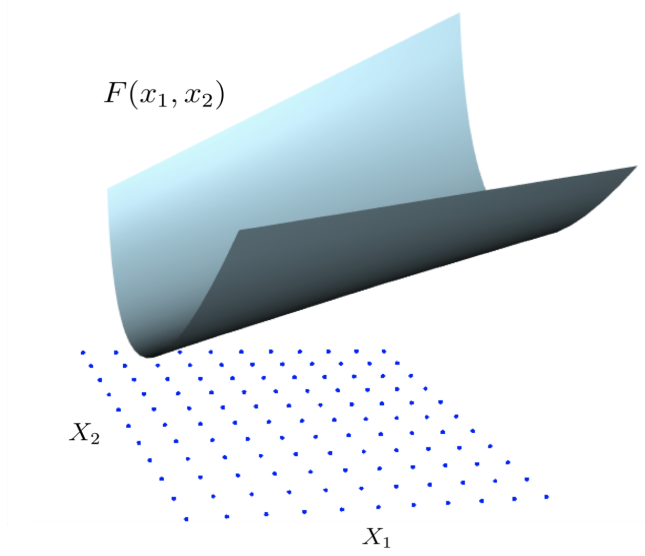


Figure 1: A grid of test points shown in the X_1X_2 plane below the response surface.

statistic to formally test the hypotheses

$$H_0 : F(x_1, x_2) = F_1(x_1) \quad \forall (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \quad (1)$$

$$H_1 : F(x_1, x_2) \neq F_1(x_1) \quad \text{for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \quad \text{for any } F_1.$$

The testing procedure we propose here follows similarly, but we impose additional structure on the test set which will allow us to perform tests for additivity and avoid training an additional set of trees.

Define a grid consisting of N total test points as in Figure 1 with N_1 levels of X_1 and N_2 levels of X_2 so that the $(i, j)^{th}$ point in the grid has true value $F_{ij} = F(X_{1_i}, X_{2_j})$ and predicted value \hat{F}_{ij} . Further, let V_F and $V_{\hat{F}}$ represent the vectorized versions of these true and predicted values so that $V_F = (F_{1,1}, \dots, F_{1,N_2}, \dots, F_{N_1,1}, \dots, F_{N_1,N_2})$ and define

$$\hat{f}_i = \frac{1}{N_2} \sum_{j=1}^{N_2} \hat{F}_{ij}$$

as the average response at the i^{th} level of X_1 across all grid levels of X_2 . For each point in the grid, define the difference in predictions $\hat{F}_{ij} - \hat{f}_i$ which we can write in vectorized form as $DV_{\hat{F}}$ for some $N \times N$ difference matrix D of rank $N - N_1$. Let Σ denote the covariance of V_F so that we can write $\Sigma_D = cov(DV_F) = D\Sigma D^T$. We then estimate Σ using the internal covariance estimation procedure proposed in Mentch and Hooker (2014) to calculate $\hat{\Sigma}$ so that $\hat{\Sigma}_D = D\hat{\Sigma}D^T$ forms a consistent estimate of Σ_D . Then, by the theoretical results established in Mentch and Hooker (2014), $(DV_{\hat{F}})^T \hat{\Sigma}_D^{-1} DV_{\hat{F}} \sim \chi_{N-N_1}^2$ and since we can equivalently write the hypotheses in (1) as

$$H_0 : F_{ij} - f_i = 0 \quad \forall (x_1, x_2) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F_{ij} - f_i \neq 0 \quad \text{for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}}$$

we can use this as a test statistic. This testing procedure is laid out in Algorithm 1. The variables $n_{\hat{\mathbf{x}}}$ and n_{MC} are covariance estimation parameters selected by the user as part of the internal covariance estimation procedure; details can be found in Mentch and Hooker (2014) and a summary is provided in Appendix A.

Algorithm 1: Test for Feature Significance

- 1 Select internal covariance estimation parameters $n_{\hat{\mathbf{x}}}$ and n_{MC}
 - 2 Perform internal covariance estimation procedure to build ensemble and estimate V_F and Σ
 - 3 Compute \hat{f}_j and calculate D and $\hat{\Sigma}_D = D\hat{\Sigma}D^T$
 - 4 Compute test statistic $(DV_{\hat{F}})^T \hat{\Sigma}_D^{-1} DV_{\hat{F}}$
-

Asymptotically, this test statistic has a $\chi_{N-N_1}^2$ distribution and thus can be compared to the $1 - \alpha$ quantile as the critical value to achieve a test with type 1 error rate α . If the test statistic is larger than the critical value, we reject the null hypothesis and conclude that X_2 is significant.

Note that this testing procedure readily extends to the more general case of d features X_1, \dots, X_d . Let \mathbf{X}_R and \mathbf{X}_A form a partition of the feature set $\{X_1, \dots, X_d\}$ so that \mathbf{X}_R and \mathbf{X}_A are disjoint and $\mathbf{X}_R \cup \mathbf{X}_A = \{X_1, \dots, X_d\}$; the set \mathbf{X}_R denotes the *reduced* set of features and \mathbf{X}_A represents the *additional* features that we want to test for significance. To test the hypotheses

$$H_0 : F(\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) = F_R(\mathbf{x}_{R_i}) \quad \forall (\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F(\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) \neq F_R(\mathbf{x}_{R_i}) \text{ for some } (\mathbf{x}_{R_i}, \mathbf{x}_{A_i}) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_R$$

we simply replace X_1 with \mathbf{X}_R and X_2 with \mathbf{X}_A in the above work by appropriately redefining levels of each feature set to form the grid of test points and repeat the procedure in Algorithm 1. Note that in this case, each grid point now corresponds to a level of a vector of features.

It is also worth noting that Mentch and Hooker (2014) suggest comparing predictions generated with the full training set not only to those produced with the reduced set \mathbf{X}_R , but also to those generated with \mathbf{X}_R and a permuted version of \mathbf{X}_A in order to rule out the possibility that the ensemble is simply making use of additional noise. The procedure we propose above avoids this problem by utilizing the projection matrix D .

3 Tests for additivity

We now demonstrate how the grid structure imposed on the test set allows for tests of additivity.

Tests for total additivity

Again assume that our training set consists of only two features X_1 and X_2 and that the response is observed according to $Y = F(X_1, X_2) + \epsilon$. Tests for total additivity assess whether the underlying regression function F is equal to, or at least well-approximated by, a sum of univariate functions so that in this 2-dimensional case, the hypotheses of interest are

$$H_0 : \exists F_1, F_2 \text{ such that } F(x_1, x_2) = F_1(x_1) + F_2(x_2) \quad \forall (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \quad (2)$$

$$H_1 : F(x_1, x_2) \neq F_1(x_1) + F_2(x_2) \text{ for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, F_2.$$

Again define a 2-dimensional grid of test points as in Figure 1 so that each point in the grid has true value F_{ij} , predicted value \hat{F}_{ij} , and vectorized versions V_F and $V_{\hat{F}}$. Define \bar{F} to be the mean of all predictions in the grid and define

$$\hat{f}_{i\cdot} = \frac{1}{N_2} \sum_{j=1}^{N_2} \hat{F}_{ij} \quad \text{and} \quad \hat{f}_{\cdot j} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{F}_{ij}$$

as the mean prediction at the i^{th} level of X_1 across all levels of X_2 , and the mean prediction at the j^{th} level of X_2 across all levels of X_1 , respectively. If the features are additive, i.e. under the null hypothesis, all points (x_{1i}, x_{2j}) in the grid can be written as $F_{ij} = f_{i\cdot} + f_{\cdot j} - \mu$ where $\mu = \mathbb{E}\bar{F}$ is the true mean expected prediction across all points in the grid. Thus, we may equivalently write the hypotheses in (2) as

$$H_0 : F_{ij} - f_{i\cdot} - f_{\cdot j} + \mu = 0 \text{ for all } (x_i, x_j) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F_{ij} - f_{i\cdot} - f_{\cdot j} + \mu \neq 0 \text{ for some } (x_1, x_2) \in \mathbf{x}_{\text{TEST}}.$$

The natural test statistic that arises is $\hat{F}_{ij} - \hat{f}_{i\cdot} - \hat{f}_{\cdot j} + \bar{F}$ which we can write in

vectorized form as $D_2 V_{\hat{F}}$ for a difference matrix D_2 . Again, let Σ denote the covariance of V_F so that we can write $\Sigma_{D_2} = \text{cov}(D_2 V_F) = D_2 \Sigma D_2^T$. Given N_1 and N_2 levels of X_1 and X_2 respectively, we estimate $P = 1 + (N_1 - 1) + (N_2 - 1)$ parameters and we can use $(D_2 V_{\hat{F}})^T \hat{\Sigma}_{D_2}^{-1} D_2 V_{\hat{F}} \sim \chi_{N-P}^2$ as our test statistic. This testing procedure for total additivity is identical to the procedure in Algorithm 1 but in Steps 3 and 4 we calculate the appropriate difference matrix and test statistic.

Note that this procedure naturally extends to the case of d features X_1, \dots, X_d . To test hypotheses of the form

$$H_0 : \exists F_1, \dots, F_d \text{ s.t. } F(x_1, \dots, x_d) = F_1(x_1) + \dots + F_d(x_d) \quad \forall (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}} \quad (3)$$

$$H_1 : F(x_1, \dots, x_d) \neq F_1(x_1) + \dots + F_d(x_d) \text{ for some } (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, \dots, F_d$$

we require a d -dimensional grid of test points so that given N_i levels of each feature X_i , our grid contains a total of $N = \prod_{i=1}^d N_i$ test points. Further, define

$$\hat{f}_{\dots j \dots} = \frac{1}{N_1 \dots N_{p-1} N_{p+1} \dots N_d} \sum_{i_1=1}^{N_1} \dots \sum_{i_{p-1}=1}^{N_{p-1}} \sum_{i_{p+1}=1}^{N_{p+1}} \dots \sum_{i_d=1}^{N_d} \hat{F}_{i_1 \dots j \dots i_d}$$

to be the average prediction over all points in the grid at the j^{th} level of X_p . As in the 2-dimensional case, we can rewrite the hypotheses in (3) as

$$H_0 : F_{i_1 \dots i_d} - f_{i_1 \dots} - f_{i_2 \dots} - \dots - f_{\dots i_d} + (d-1)\mu = 0 \text{ for all } (x_{i_1}, \dots, x_{i_d}) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F_{i_1 \dots i_d} - f_{i_1 \dots} - f_{i_2 \dots} - \dots - f_{\dots i_d} + (d-1)\mu \neq 0 \text{ for some } (x_{i_1}, \dots, x_{i_d}) \in \mathbf{x}_{\text{TEST}}$$

and write $\hat{F}_{i_1 \dots i_d} - \hat{f}_{i_1 \dots} - \dots - \hat{f}_{\dots i_d} + (d-1)\bar{F}$ as $D_d V_{\hat{F}}$. Again, we define Σ to be the covariance of V_F so that $\Sigma_{D_d} = \text{cov}(D_d V_F) = D_d \Sigma D_d^T$ and we can use $(D_d V_{\hat{F}})^T \hat{\Sigma}_{D_d}^{-1} D_d V_{\hat{F}} \sim \chi_{N-P}^2$ as our test statistic, where $P = 1 + (N_1 - 1) + \dots + (N_d - 1)$.

Furthermore, the additive functions need not be univariate. Define a (disjoint) partition of the feature space $\mathbf{S}_1, \dots, \mathbf{S}_m$ so that $\cup_{i=1}^m \mathbf{S}_i = \{X_1, \dots, X_d\}$. We can test hypotheses of the form

$$H_0 : \exists F_1, \dots, F_m \text{ such that } F(\mathbf{s}_1, \dots, \mathbf{s}_m) = F_1(\mathbf{s}_1) + \dots + F_m(\mathbf{s}_m) \quad \forall (\mathbf{s}_1, \dots, \mathbf{s}_m) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F(\mathbf{s}_1, \dots, \mathbf{s}_m) \neq F_1(\mathbf{s}_1) + \dots + F_m(\mathbf{s}_m) \text{ for some } (\mathbf{s}_1, \dots, \mathbf{s}_m) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, \dots, F_m$$

in exactly the same fashion by appropriately defining levels of an m -dimensional grid of test points.

Tests for partial additivity

We now handle the case where we are interested in testing only whether a proper subset of features contribute additively to the response. Suppose that our training set consists of three features X_1, X_2 and X_3 and we are interested in testing

$$H_0 : \exists F_1, F_2 \text{ s.t. } F(x_1, x_2, x_3) = F_1(x_1, x_3) + F_2(x_2, x_3) \quad \forall (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}} \quad (4)$$

$$H_1 : F(x_1, x_2, x_3) \neq F_1(x_1, x_3) + F_2(x_2, x_3) \text{ for some } (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_1, F_2.$$

Rejecting this null hypothesis means that an interaction exists between X_1 and X_2 but implies nothing about potential interactions between X_1 and X_3 or between X_2 and X_3 . Hooker (2004) uses the size of the deviation of F from partial additivity as a means of identifying the bivariate and higher-order interactions required to reconstruct some percentage of the variation in the values of F . This is also referred to as the Sobol index for the X_1, X_2 interaction (Sobol (2001)). Define a 3-dimensional grid of test points with N_1, N_2 and N_3 levels of X_1, X_2 and X_3 , respectively and continuing with the dot notation, define

$$\hat{f}_{i \cdot k} = \frac{1}{N_2} \sum_{j=1}^{N_2} \hat{F}_{ijk} \quad \text{and} \quad \hat{f}_{\cdot jk} = \frac{1}{N_1} \sum_{i=1}^{N_1} \hat{F}_{ijk}$$

to be the average prediction over all levels of X_2 in the grid at the i^{th} and k^{th} levels of X_1 and X_3 , and the average prediction over all levels of X_1 in the grid at the j^{th} and k^{th} levels of X_2 and X_3 , respectively. If there is no interaction between X_1 and X_2 , then $F_{ijk} - f_{i \cdot k} - f_{\cdot jk} + f_{\cdot \cdot k} = 0$ at all levels $(x_{1_i}, x_{2_j}, x_{3_k})$ in the grid. Thus, we can rewrite the hypotheses in (4) as

$$H_0 : F_{ijk} - f_{i \cdot k} - f_{\cdot jk} + f_{\cdot \cdot k} = 0 \quad \forall (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F_{ijk} - f_{i \cdot k} - f_{\cdot jk} + f_{\cdot \cdot k} \neq 0 \quad \text{for some } (x_1, x_2, x_3) \in \mathbf{x}_{\text{TEST}}$$

and use the empirical analogues of these parameters to conduct the testing procedure. Once again, write $\hat{F}_{ijk} - \hat{f}_{i \cdot k} - \hat{f}_{\cdot jk} + \hat{f}_{\cdot \cdot k}$ as $D_3 V_{\hat{F}}$ and define Σ to be the covariance of V_F so that we can write $\Sigma_{D_3} = \text{cov}(D_3 V_F) = D_3 \Sigma D_3^T$ and use $(D_3 V_{\hat{F}})^T \hat{\Sigma}_{D_3}^{-1} D_3 V_{\hat{F}} \sim \chi_{N-P}^2$ as our test statistic, where $N = N_1 N_2 N_3$ and since we must now account for two-way interactions, $P = 1 + (N_1 - 1) + (N_2 - 1) + (N_3 - 1) + (N_1 - 1)(N_3 - 1) + (N_2 - 1)(N_3 - 1)$. As was the case in testing for total additivity, the testing procedure is the same as in Algorithm 1 with the appropriate difference matrix and test statistic calculated in Steps 3 and 4.

Note that we can perform this same test when our training set consists of d features and we are interested in determining whether an interaction exists between X_i and X_j . Denote the set of all features except X_i and X_j as $\mathbf{X}_{-i,j}$ so that our hypotheses

become

$$H_0 : \exists F_i, F_j \text{ such that } F(X_1, \dots, X_d) = F_i(X_i, \mathbf{X}_{-i,j}) + F_j(X_j, \mathbf{X}_{-i,j}) \quad \forall (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}}$$

$$H_1 : F(X_1, \dots, X_d) \neq F_i(X_i, \mathbf{X}_{-i,j}) + F_j(X_j, \mathbf{X}_{-i,j}) \text{ for some } (x_1, \dots, x_d) \in \mathbf{x}_{\text{TEST}} \text{ for any } F_i, F_j.$$

By simply redefining the grid levels of X_3 to be levels of $\mathbf{X}_{-i,j}$, the testing procedure remains identical. The same holds for the case where X_i and X_j are treated as vectors.

A general approach

The above testing procedures were derived by choosing the model parameters that minimized the sum of squared error (SSE) with equal weight placed on each point in the test grid. We could have instead selected \hat{F}_1 and \hat{F}_2 to minimize the weighted SSE

$$WSSE = \sum_{i,j,k} w_{i,j,k} \left(F(x_{1_i}, x_{2_j}, x_{3_k}) - F_1(x_{1_i}, x_{3_k}) - F_2(x_{2_j}, x_{3_k}) \right)^2$$

where the w_{ijk} are specified by the user. Hooker (2007), for example, recommends basing these weights on an approximation to the density of observations near the point $(x_{1_i}, x_{2_j}, x_{3_k})$. This procedure takes the form of a weighted ANOVA. In particular, define \vec{F} to be the $N_1 N_3 + N_2 N_3$ vector concatenating the $\hat{F}_1(x_1, x_3)$ and $\hat{F}_2(x_2, x_3)$ and as in the previous sections let $V_{\hat{F}}$ be the vector containing the \hat{F}_{ijk} . Further, let Z be the $N \times N_1 N_3 + N_2 N_3$ matrix defined so that $Z\vec{F}$ produces the corresponding $\hat{F}_1(x_1, x_3) + \hat{F}_2(x_2, x_3)$ and let W be a diagonal matrix containing the weights. Then we can write

$$WSSE = (V_{\hat{F}} - Z\vec{F})^T W (V_{\hat{F}} - Z\vec{F})$$

and we know that the solution \vec{F} that minimizes this weighted SSE is given by

$$\vec{F} = (Z^T W Z)^{-1} Z^T W V_{\hat{F}}$$

so that under the null hypothesis

$$V_{\hat{F}} - Z\vec{F} = (I - Z(Z^T W Z)^{-1} Z^T W) V_{\hat{F}}$$

has mean 0. Further, letting Σ denote the covariance of V_F , the variance of $V_F - Z\vec{F}$ is given by

$$C = (I - Z(Z^T W Z)^{-1} Z^T W) \Sigma (I - Z(Z^T W Z)^{-1} Z^T W)^T$$

so that

$$[(I - Z(Z^T W Z)^{-1} Z^T W) V_{\hat{F}}]^T \hat{C}^{-1} [(I - Z(Z^T W Z)^{-1} Z^T W) V_{\hat{F}}]$$

has a χ^2_{N-P} distribution. For equal weighting (W given by the identity matrix), these calculations reduce to the averages employed above, and for the sake of simplicity we have restricted ourselves to this choice in the examples below.

4 Extension to Large Grids

Note that the above procedures require estimating a covariance matrix of size proportional to the number of points in the test grid. In our experience, using relatively few trees to estimate the covariance parameters usually leads to an overestimate, which

can thereby reduce the power in the testing procedures. Thus, in situations where many grid points are required, either because certain features contain many levels or because we are interested in evaluating a more complex additive structure, it may be computationally infeasible to directly obtain a reasonable estimate. To account for this potential large grid problem, we borrow from recently developed methods utilizing random projections.

4.1 Random Projections

We present here a brief overview of random projection theory; for more background we refer the reader to Bingham and Mannila (2001). Let $X_{n \times p}$ represent a data matrix with sample size n and presumably large number of covariates p , perhaps even $p > n$. The idea is to project the data into a lower dimensional subspace via a random projection matrix $R_{p \times r}$ where $1 \leq r < n$ to form the projected data matrix $XR_{n \times r}$. The Lemma provided by Johnson and Lindenstrauss (1984) states that an orthogonal projection of elements in a higher dimensional space into a lower dimensional space will approximately preserve the distances between projected elements. Thus, selecting an orthogonal random projection matrix R with relatively small projected dimension r allows us to take initially high dimensional data and work in a more friendly lower dimensional space without a significant distortion of information.

Recently, Srivastava et al. (2014) proposed a new high dimensional testing procedure called RAPTT (**R**andom **P**rojection **T**-Test) that represents an extension of Hotelling's classic T^2 test to the $p > n$ case. The authors utilize random projections to test equality of multivariate means and demonstrate that this procedure performs well with reasonably high power even in a $p \gg n$ setting. Given two multivariate samples $X_{n_1 \times p}$ and $Y_{n_2 \times p}$, the test statistic for the projected data is given by

$$T_R^2 = \frac{1}{n_1^{-1} + n_2^{-1}} (\bar{X} - \bar{Y})' R (R' S R)^{-1} R' (\bar{X} - \bar{Y})$$

and the authors obtain a randomized test by defining

$$\phi(T_R^2) = \begin{cases} 1 & \text{if } \frac{n-r+1}{r} \frac{T_R^2}{n} > c_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where $n = n_1 + n_2 - 2$ and c_α is taken as the $1 - \alpha$ quantile of the $F_{r, n-r+1}$ distribution. Though this test is unbiased and consistent of exact size α , it may have little power and lead to different conclusions based on the choice of R . RAPTT is thus defined by averaging the p-values across M such tests, utilizing M random projections R_1, \dots, R_M . The p-value of the i^{th} such test is given by

$$\theta_i = 1 - F_{r, n-r+1} \left(\frac{n-r+1}{r} \frac{T_R^2}{n} \right)$$

and RAPTT is defined by

$$\bar{\phi} = \begin{cases} 1 & \text{if } \bar{\theta} < u_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{\theta} = \frac{1}{M} \sum_{i=1}^M \theta_i$ and u_α is chosen such that $P[\bar{\theta} < u_\alpha | H_0] = \alpha$.

4.2 Application to Ensemble Tests

We now demonstrate how this idea of testing using random projections can be applied to our procedures. Suppose we have a training set of size n and we build an ensemble consisting of m trees, each of which is built with a subsample of size k , and that we are interested in predicting at N total test points. The test statistic for the original

test of feature significance proposed in Mentch and Hooker (2014) can be written as $\hat{\mu}_N^T \hat{\Sigma}^{-1} \hat{\mu}_N$ where $\hat{\mu}$ is the vector of ensemble predictions, and $\hat{\Sigma}$ is a consistent estimate of the corresponding covariance matrix. Based on the theory in Mentch and Hooker (2014), this test statistic follows a χ_N^2 distribution and we can employ a testing procedure based on random projections similar to the RAPTT procedure outlined above. Given m predictions at each of N locations, we can think of our data as an $m \times N$ matrix so that for a set of random projection matrices R_1, \dots, R_M and reduced dimension $r < m, N$, we can write the projected test statistic as

$$T_R = \hat{\mu}_N^T R (R^T \hat{\Sigma} R)^{-1} R^T \hat{\mu}_N \sim \chi_r^2. \quad (5)$$

In the case where we perform the internal variance estimation procedure to simultaneously build the ensemble and estimate the variance parameters, we can think of our data matrix as having dimension $n_{\tilde{x}} \times N$, in which case we choose $r < n_{\tilde{x}}, N$. In the context of testing for additivity via the grid approach outlined above, the procedure remains largely the same. We are already utilizing a difference matrix D to project into the space of additive models, but so long as the elements of R are independently generated continuous random variables, the overall projection has rank r with probability 1. Recall that the original test statistic is given by $(DV_{\hat{F}})^T \hat{\Sigma}_D^{-1} DV_{\hat{F}}$ where $\Sigma_D = \text{cov}(DV_{\hat{F}}) = D \Sigma D^T$ and so our random projection test statistic becomes

$$T_R = (DV_{\hat{F}})^T R (R^T \hat{\Sigma}_D R)^{-1} R^T (DV_{\hat{F}}) \sim \chi_r^2$$

where now, $r < N - P$. In both cases, the p-value for a single implementation of our random projection test based on the projection matrix R_i is given by

$$\theta_i = 1 - \Phi_r^2(T_R)$$

where Φ_r^2 denotes the cdf of the χ_r^2 . Thus, for M repetitions of this test, we can define our final test in exactly the same fashion as RAPTT, so that

$$\bar{\phi} = \begin{cases} 1 & \text{if } \bar{\theta} < u_\alpha \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{\theta} = \frac{1}{M} \sum_{i=1}^M \theta_i$ and u_α is chosen such that $P[\bar{\theta} < u_\alpha | H_0] = \alpha$.

4.3 Choice of R and r

Srivastava et al. (2014) note that the method of sampling projection matrices R_i and the choice of reduced dimension r are important considerations in this procedure as the power of the test can depend on both. In our case, the covariance matrix parameters Σ_{1,k_n} and Σ_{k_n,k_n} are usually difficult to accurately estimate when a large number of grid points is required. Thus, though the procedure is well defined for $1 \leq r < m$, this practical restriction necessitates a relatively small projected dimension r . In our experience, we see a significant drop in power whenever our grids consist of more than approximately 30 points, so choosing $5 \leq r \leq 15$ should be reasonable and computationally feasible in most situations.

Note also that because r is small, little dependence remains between the resulting p-values θ_i . Under the null hypothesis, each p-value is uniformly distributed on $[0, 1]$. Furthermore, the mean of independent standard uniform random variables follows a Bates distribution, so the final cutoff u_α can be well approximated by the α quantile of this distribution when r is small. Srivastava et al. (2014) suggest simulating under the null hypothesis to estimate u_α but in our case, these simulations are difficult to perform since we cannot assume that the rows in our data matrix come from a multivariate normal distribution.

The most appropriate choice of random projection matrices is less clear. Srivastava et al. (2014) demonstrate that any semi-orthogonal matrix R (i.e. $R^T R = I_r$, where I_r is the $r \times r$ identity matrix) whose elements are continuous random variables with finite second moment will satisfy the necessary conditions to perform the tests based on random projections but consider both a uniform sample from the Haar distribution and an idea referred to as *one permutation + one random projection* in developing RAPTT. Bingham and Mannila (2001) note that the elements of R are typically sampled from a normal distribution, but that they may be selected from simpler distributions, such as that suggested by Achlioptas (2001).

Algorithm 2: Internal Estimation Procedure

- 1 Compute difference matrix D
 - 2 Select reduced dimension r
 - 3 Generate random projection matrices R_1, \dots, R_M
 - 4 **For** i in 1 to $n_{\tilde{\mathbf{x}}}$
 - 5 Select initial fixed point $\tilde{\mathbf{x}}^{(i)}$
 - 6 **For** j in 1 to n_{MC}
 - 7 Select subsample $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)},j}$ of size k_n from training set that includes $\tilde{\mathbf{x}}^{(i)}$
 - 8 Build tree using subsample $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)},j}$
 - 9 Use tree to predict at each of the N grid points to obtain \hat{V}_j
 - 10 Apply each projection to \hat{V} to obtain $\hat{W}_{j,c} = (D\hat{V}_j)^T R_c$
 - 11 **End**
 - 12 Record average of $\hat{W}_{j,c}$ over j for each projection
 - 13 **End**
 - 14 Compute variance of each of the $n_{\tilde{\mathbf{x}}}$ averages to estimate each ζ_{1,k_n}
 - 15 Compute variance of all predictions from each projection to estimate each ζ_{k_n,k_n}
 - 16 Compute mean of all predictions from each projection to estimate each θ_{k_n}
 - 17 Compute each p-value $\theta_1, \dots, \theta_M$ by comparing to χ_r^2
 - 18 Record average p-value $\bar{\theta}$ and compare to Bates α quantile
-

For our situation, we recommend simply generating random matrices whose elements are sampled from a standard normal, orthogonalizing via a process such as Gram-Schmidt, and selecting the appropriate submatrix. Though such a method is

perhaps not the most computationally efficient – as noted by Bingham and Manilla (2001) – it is straightforward and can be easily implemented in most software packages.

Our testing procedure based on random projections is laid out in Algorithm 2. It is worth emphasizing again that though the original ensemble does not need to be built according to the internal estimation procedure, building in this fashion allows us to more easily use a small projected dimension r and also allows for simultaneous estimation of the variance parameters. It is also worth noting that the classic sample covariance S is consistent and could potentially be used in place of the variance estimation procedures suggested by Mentch and Hooker (2014), but typically does not produce an accurate estimate.

5 Simulations

We now provide a short simulation study to investigate both the α -level and power of our proposed testing procedures. Suppose first that we have two features X_1 and X_2 and that our responses are generated according to

$$Y = X_1 + X_2 + \beta X_1 X_2 + \epsilon$$

where $\beta = 0$ in assessing the α -level and $\beta = 1$ to evaluate power and $\epsilon \sim \mathcal{N}(0, 0.05^2)$. We repeat the test for total additivity on 1000 datasets when $\beta = 0$ and 1000 where $\beta = 1$, taking our empirical α -level as the proportion of tests that incorrectly reject the null hypothesis (when $\beta = 0$) and our estimate of power as the proportion of tests that correctly reject the null hypothesis (when $\beta = 1$). For reference, we also built 1000 linear regression models and used the traditional t-test to determine whether

Method	n	α -level	Power
Linear Model	250	0.056	1.000
Subbagged Ensemble		0.065	0.954
Linear Model	500	0.048	1.000
Subbagged Ensemble		0.047	0.998
Linear Model	1000	0.046	1.000
Subbagged Ensemble		0.020	0.999

Table 1: Empirical α -levels and power for the linear model example.

the interaction coefficient is significant on each dataset and recorded the empirical α -level and power of this testing procedure. This was repeated for training set sizes of 250, 500, and 1000 using subsample sizes of 30, 50, and 75 respectively and the results are shown in Table 1. The test grid was selected as a 4×4 grid with levels 0.2, 0.4, 0.6, and 0.8. In each case, our test for total additivity using a subbagged ensemble performed nearly exactly as well as the traditional t-test.

We also selected a number of more complex regression functions to further investigate the α -level and power. Many of these underlying regression functions were studied in previous publications relating to tests for additivity, such as De Canditiis and Sapatinas (2004) and Barry (1993). Each estimate is the result of 1000 simulations with a sample size of 500, subsample size of 50, and internal covariance estimation parameters $n_{\tilde{\mathbf{x}}} = 50$ and $n_{MC} = 250$ with a 4×4 test grid (with levels 0.2, 0.4, 0.6, and 0.8) in the 2-dimensional tests for total additivity and a $3 \times 3 \times 3$ grid (with levels 0.3, 0.5, and 0.7) in the 3-dimensional tests for total and partial additivity. In each case the features were selected uniformly at random from $[0, 1]$ and the responses were generated according to $Y = F(\mathbf{X}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.05^2)$. The results are shown in Table 2. Note that even though the response in the first two models does not depend on X_2 , this additional feature was still included in the training sets and the same test for total additivity was performed. In each case, we

Model	Test	α -level	Model	Test	Power
$y = x_1$	Total	0.000	$y = x_1x_2$	Total	1.000
$y = e^{x_1}$	Total	0.000	$y = x_1x_2x_3$	Partial	0.605
$y = e^{x_1} + \sin(\pi x_2)$	Total	0.004	$y = e^{5(x_1+x_2)} / (1 + e^{5(x_1+x_2)}) - 1$	Total	0.948
$y = x_1 + x_2 + x_3$	Total	0.002	$y = 0.5(1 + \sin(2\pi(x_1 + x_2)))$	Total	1.000
$y = e^{x_1} + e^{x_2} + e^{x_3}$	Total	0.004	$y = 0.5(1 + \sin(2\pi(x_1 + x_2 + x_3)))$	Partial	0.778
$y = x_1x_3 + x_2x_3$	Partial	0.000	$y = 64(x_1x_2)^3(1 - x_1x_2)^3$	Total	1.000
$y = e^{x_1x_3} + e^{x_2x_3}$	Partial	0.000	$y = 64(x_1x_2x_3)^3(1 - x_1x_2x_3)^3$	Partial	0.956

Table 2: Empirical α -level and power for a variety of underlying regression functions.

see that our false rejection rate is very conservative and we also maintain high power.

Finally, we repeated these simulations on the same regression functions, this time employing our testing procedures based on random projections. In the 2-dimensional tests for total additivity, we use a 10×10 test grid so that $N = 100$ and in the 3-dimensional tests for total additivity and the tests for partial additivity, we use a $5 \times 5 \times 5$ test grid for a total of 125 test points. The results are shown in Table 3. Note that in these tests, we maintain a reasonable type 1 error rate but have significantly more power due to the finer resolution of the test grid.

Model	Test	α -level	Model	Test	Power
$y = x_1$	Total	0.000	$y = x_1x_2$	Total	1.000
$y = e^{x_1}$	Total	0.000	$y = x_1x_2x_3$	Partial	0.998
$y = e^{x_1} + \sin(\pi x_2)$	Total	0.059	$y = e^{5(x_1+x_2)} / (1 + e^{5(x_1+x_2)}) - 1$	Total	0.999
$y = x_1 + x_2 + x_3$	Total	0.001	$y = 0.5(1 + \sin(2\pi(x_1 + x_2)))$	Total	1.000
$y = e^{x_1} + e^{x_2} + e^{x_3}$	Total	0.012	$y = 0.5(1 + \sin(2\pi(x_1 + x_2 + x_3)))$	Partial	0.959
$y = x_1x_3 + x_2x_3$	Partial	0.008	$y = 64(x_1x_2)^3(1 - x_1x_2)^3$	Total	1.000
$y = e^{x_1x_3} + e^{x_2x_3}$	Partial	0.066	$y = 64(x_1x_2x_3)^3(1 - x_1x_2x_3)^3$	Partial	1.000

Table 3: Empirical α -level and power for the testing procedure based on random projections.

6 Real data

We now demonstrate our testing procedures on a dataset provided by a team of ornithologists at the Cornell University Lab of Ornithology as part of the *ebird* project described in Sullivan et al. (2009). This ongoing project relies on citizens to submit reports of bird observations and from these reports, researchers are able to monitor things like migration patterns and species abundance. The dataset we were provided was originally compiled in order to determine how pollution levels affect the change in Wood Thrush population. The data consists of 3 pollutant features, mercury deposition (*md*), acid deposition (*ad*), and soil PH level (*sph*) as well as 2 non-pollutant features, elevation (*elev*) and abundance (*ab*). We begin our analysis by testing whether the pollutant and non-pollutant features are additive:

$$H_0 : F(md, ad, sph, elev, ab) = F_P(md, ad, sph) + F_{NP}(elev, ab). \quad (6)$$

We performed a test for total additivity using 4 levels of each feature set, the 0.20, 0.40, 0.60, and 0.80 quantiles of each feature, for a total of 16 test points. Our test statistic was 52.30, larger than the critical value, the 0.95 quantile of the χ^2_9 , of 16.92 so we reject the null hypothesis in (6) and conclude that an interaction exists between the pollutant and non-pollutant features. This result was confirmed by our random projection test, which consisted of 1000 random projections to a dimension of $r = 5$ using a 10×10 test grid. In this case, the final averaged p-value was only 0.0043, far below the critical value of 0.485.

Next, we investigated how the pollutants contributed to the response. Based on preliminary investigations, ebird researchers suspected an interaction between mercury and acid deposition (*md* and *ad*) but were unsure of the relationship between

soil PH (*sph*) and *md* and *ad*. In performing these tests for partial additivity, our test grid consisted of 3 points for each feature set, the 0.30, 0.50, and 0.70 quantiles of each feature for a total of 27 test points and a critical value, the 0.95 quantile of the χ^2_{12} , of 21.03. Our test for partial additivity between *md* and *ad*,

$$H_0 : F(md, ad, sph, elev, ab) = F_1(md, sph, elev, ab) + F_2(ad, sph, elev, ab).$$

yielded a significant result with a test statistic of 41.00 so our test supports the belief that an interaction exists between *md* and *ad*. Again, this result was supported by our random projection test, which consisted of 1000 random projections to a dimension of $r = 5$ using a $5 \times 5 \times 5$ test grid, for a total of 125 test points. The final averaged p-value was only 0.0064, far below the critical value of 0.485.

Our test for partial additivity between *sph* and the vector (md, ad)

$$H_0 : F(md, ad, sph, elev, ab) = F_1(md, ad, elev, ab) + F_2(sph, elev, ab).$$

yielded a test statistic of 36.43, above the critical value of 21.03, so once again we reject the null hypothesis and conclude that an interaction exists between *sph* and (md, ad) . This result was again supported by the random projection test based on 1000 random projections to a dimension of $r = 5$ using a $5 \times 5 \times 5$ test grid. We find a final averaged p-value of 0.225, which, though larger than in the previous tests, is still far below the critical value of 0.485.

7 Discussion

This work continues the recent trend of developing formal statistical inference procedures within the context of learning algorithms. The desirable asymptotic properties

of subsampled ensembles explored by Mentch and Hooker (2014) and Wager (2014) have begun to shed light on the black-box learning algorithms and the tests for additivity developed here further demonstrate that traditional scientific and statistical questions need not be seen as a sacrifice of less interpretable machine learning procedures.

An important aspect of our testing procedures is the definition of the test grid. In our applications we chose a relatively small grid in order achieve reasonably accurate covariance estimates while still maintaining an acceptable amount of power. When larger testing grids are required, we suggest utilizing random projections in order to maintain high power. Selecting levels of each feature away from the boundary of the feature space is also important as trees are known to exhibit bias near the edge of the feature space.

In the preceding work, all of our tests were carried out using subbagged ensembles, but the theory established in Mentch and Hooker (2014) and Wager (2014) also allows for subsampled random forests or any other ensemble learner conforming to the regularity conditions. However, the predictive improvement often seen with random forests is generally attributed to the increased independence between trees in the ensemble. Since our base learners are built with subsamples instead of full bootstrap samples and thus are less dependent, we expect the predictive improvement often seen with random forests to be less significant.

It is also worth noting that the particular additive forms for which we developed testing procedures were selected only because of their scientific utility. Other similar testing procedures could also be developed in the same manner by establishing the appropriate model parameters from an ANOVA set-up and defining the difference matrix D accordingly. These methods can also be extended to provide formal statistical guarantees for the screening procedures described in Hooker (2004).

References

- Achlioptas, D. (2001). Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM.
- Amato, U. and Antoniadis, A. (2001). Adaptive wavelet series estimation in separable nonparametric regression models. *Statistics and Computing*, 11(4):373–394.
- Barry, D. (1993). Testing for additivity of a regression function. *The Annals of Statistics*, pages 235–254.
- Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.
- Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.
- De Canditiis, D. and Sapatinas, T. (2004). Testing for additivity and joint effects in multivariate nonparametric regression using fourier and wavelet methods. *Statistics and Computing*, 14(3):235–249.
- Derbort, S., Dette, H., and Munk, A. (2002). A test for additivity in nonparametric regression. *Annals of the Institute of Statistical Mathematics*, 54(1):60–82.

- Dette, H. and Derbort, S. (2001). Analysis of variance in nonparametric regression models. *Journal of multivariate analysis*, 76(1):110–137.
- Dette, H., Wilkau, C. V. L. U., et al. (2001). Testing additivity by kernel-based methods-what is a reasonable test? *Bernoulli*, 7(4):669–697.
- Eubank, R., Hart, J. D., Simpson, D., Stefanski, L. A., et al. (1995). Testing for additivity in nonparametric regression. *The Annals of Statistics*, 23(6):1896–1920.
- Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association*, 100(471):890–907.
- Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43. CRC Press.
- Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580. ACM.
- Hooker, G. (2007). Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3).
- Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1.
- Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pages 93–100.

- Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM.
- Mammen, E., Linton, O., Nielsen, J., et al. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27(5):1443–1490.
- Mentch, L. and Hooker, G. (2014). Ensemble Trees and CLTs: Statistical Inference for Supervised Learning. arXiv:1404.6473 [stat.ML].
- Opsomer, J. D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, 93(442):605–619.
- Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, 8(4):715–732.
- Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.
- Srivastava, R., Li, P., and Ruppert, D. (2014). RAPTT: An Exact Two-Sample Test in High Dimensions Using Random Projections. arXiv:1405.1792 [stat.ME].
- Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., and Kelling, S. (2009). ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292.

Wager, S. (2014). Asymptotic Theory for Random Forests. arXiv:1405.0352 [math.ST].

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651.

Appendix

A Internal Estimation Procedure

We provide here a review of the internal estimation procedure suggested in Mentch and Hooker (2014). We begin by providing one of the main results of this work which establishes asymptotic normality and allows for formalized inference.

Theorem 1. Let $X_1, X_2, \dots \stackrel{iid}{\sim} F_X$ and let U_{n,k_n,m_n} be an incomplete, infinite order U -statistic with kernel h_{k_n} and $\theta_{k_n} = \mathbb{E}h_{k_n}(X_1, \dots, X_{k_n})$ such that $\mathbb{E}h_{k_n}^2(X_1, \dots, X_{k_n}) < \infty$ for all n . Suppose that $\lim \frac{k_n}{\sqrt{n}} = 0$ and let $\lim \frac{n}{m_n} = \alpha$. Then as long as $\lim k_n^2 \zeta_{1,k_n} \neq 0$,

- (i) if $\alpha = 0$, then $\frac{\sqrt{n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{k_n^2 \zeta_{1,k_n}}} \xrightarrow{d} \mathcal{N}(0, 1)$.
- (ii) if $0 < \alpha < \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\frac{k_n^2}{\alpha} \zeta_{1,k_n} + \zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0, 1)$.
- (iii) if $\alpha = \infty$, then $\frac{\sqrt{m_n}(U_{n,k_n,m_n} - \theta_{k_n})}{\sqrt{\zeta_{k_n,k_n}}} \xrightarrow{d} \mathcal{N}(0, 1)$.

Thus, provided that the subsample size k is on the order of \sqrt{n} , the predictions follow one of the three limiting distributions given above, and the limiting variance depends only on the ratio of the training set size to the number of base learners in the ensemble. The center of the distribution is estimated with the standard (point)

prediction generated by the ensemble and the parameter α can be approximated by simply plugging in n/m_n but the variance parameters ζ_{1,k_n} and ζ_{k_n,k_n} (or the multivariate analogues Σ_{1,k_n} and Σ_{k_n,k_n} used in the testing procedures) still need to be estimated. For a general c , these variance parameters are defined as

$$\zeta_{c,k_n} = \text{var}\left(\mathbb{E}(h_{k_n}(X_1, \dots, X_{k_n}) \mid X_1 = x_1, \dots, X_c = x_c)\right)$$

so that

$$\zeta_{1,k_n} = \text{var}\left(\mathbb{E}(h_{k_n}(X_1, \dots, X_{k_n}) \mid X_1 = x_1)\right)$$

and

$$\zeta_{k_n,k_n} = \text{var}(h_{k_n}(X_1, \dots, X_{k_n})).$$

The procedure to estimate ζ_{1,k_n} is provided below. The procedure to estimate Σ_{1,k_n} is nearly identical, but use the ensemble to predict at all points in step 6, and instead of taking the variance of the vector of $n_{\tilde{\mathbf{x}}}$ averages, we record the covariance of $n_{\tilde{\mathbf{x}}}$ rows of averages.

Algorithm 3: ζ_{1,k_n} Estimation Procedure

- 1 **For** i in 1 to $n_{\tilde{\mathbf{x}}}$
 - 2 Select initial fixed point $\tilde{\mathbf{x}}^{(i)}$
 - 3 **For** j in 1 to n_{MC}
 - 4 Select subsample $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)},j}$ of size k_n from training set that includes $\tilde{\mathbf{x}}^{(i)}$
 - 5 Build tree using subsample $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)},j}$
 - 6 Use tree to predict at \mathbf{x}^*
 - 7 **End**
 - 8 Record average of the n_{MC} predictions
 - 9 **End**
 - 10 Compute the variance of the $n_{\tilde{\mathbf{x}}}$ averages
-

To estimate ζ_{k_n, k_n} we can simply generate a large number of subsamples and record the variance (or covariance matrix) of the predictions. Alternatively, to be more computationally efficient, we can record all predictions generated in estimating ζ_{1, k_n} and record the variance (covariance) amongst these as our estimate.

Algorithm 4: Internal Estimation Procedure

```

1  For  $i$  in 1 to  $n_{\tilde{\mathbf{x}}}$ 
2      Select initial fixed point  $\tilde{\mathbf{x}}^{(i)}$ 
3      For  $j$  in 1 to  $n_{MC}$ 
4          Select subsample  $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)}, j}$  of size  $k_n$  from training set that includes  $\tilde{\mathbf{x}}^{(i)}$ 
5          Build tree using subsample  $\mathcal{S}_{\tilde{\mathbf{x}}^{(i)}, j}$ 
6          Use tree to predict at  $\mathbf{x}^*$  and record prediction
7      End
8      Record average of the  $n_{MC}$  predictions
9  End
10 Compute the variance of the  $n_{\tilde{\mathbf{x}}}$  averages to estimate  $\zeta_{1, k_n}$ 
11 Compute the variance of all predictions to estimate  $\zeta_{k_n, k_n}$ 
12 Compute the mean of all predictions to estimate  $\theta_{k_n}$ 

```

These variance estimation procedures can be performed after the original ensemble is computed, which is referred to as *external* estimation. However, this means that all of the predictions generated in an effort to estimate ζ_{1, k_n} simply go to waste. To perform the *internal* variance estimation procedure, we record all predictions used in Algorithm 3, using the average of all predictions as our point estimate and the variance across all predictions as our estimate of ζ_{k_n, k_n} . The algorithm to perform this internal estimation is provided in Algorithm 4.